Semeion

5   An Algorithm for projecting information data belonging
    to a multidimensional space into a space having less
    dimensions a method for the cognitive analysis of
    multidimensional information data based on the said
    algorithm and a program comprising the said algorithm
10 stored on a recordable support and an apparatus having
    artificial intelligence.

    The invention relates to an algorithm for
    projecting information data belonging to a
15 multidimensional space into a space having less
    dimensions.

    The invention relates particularly to the field of
    artificial intelligence and the aim is to allow a
    machine able to carry out computational tasks to
20 analyse complex n-dimensional data in order to
    represent this data in a two or three dimensional space
    and so to evaluate this data for cognitive tasks, as
    for example to create a simplified and representable
    image of the data or to evaluate the existence of
25 relationships between a group of data records which
    relationships cannot be represented by exact computable
    or mathematical functions or for computational tasks in
    order to solve a problem which is not based on exact
    mathematical functions.

30   As it is known nature cannot be always represented
    by functions having an exact solution or by system of
    equations having a tall a mathematical solution. In the
    exact sciences a model may be constructed for

simplifying the relationships and helping the mathematical inspection to be carried out in order to achieve a mathematical representation of the relation ships between data or a correlation among data

5   appearing as not correlated and a mathematical tool for evaluating the degree or level of the correlation of the data. Furthermore the models may consist in recognizing and constructing images or structures and provide for a graphic representation.

10   Further to the problem of having tools for apparati having artificial intelligence in order to better understand, classify and evaluate the physical or chemical world and nature, it has to be noted that artificial intelligence is not limited to the analysis

15   and inspection of nature only relatively to exact scientific or technical problems or structures but must be also confronted with social problems which are far most difficult to be represented by the mathematical tools or by exact computable functions. In this case

20   the apparatus is confronted with individuals having a specific behaviour and acting on their own mind or by means of reactions to instincts which actions cannot be described by mathematical models because there is no mathematical model and also because there is no clear

25   and univoque rule defining the relations between events whichever kind they are and the behaviour.

            Human beings has the capacity of analyzing environmental stimuli and deciding to carry out an action as a response to the said stimuli also when

30   apparently the stimuli have no relationship among them or are not correlated. This process is carried out sometimes in a non conscious way giving rise to logically non predictable actions if considering the

known relationship of the stimuli if one ever exists. Nevertheless the action is often correct or approximately correct or leads to a certain successful effect. Such kind of behaviour which we can define as intuition or the like seems not to have any logical basis or seems not to be caused by a logical thought.

Since artificial intelligence is based on computational machines there is the need of instruments which may help this machines to analyzed or transform information data in such a way as to be simply handled and used by the machine and in such a way as to allow the machine to recognize and/or generate relationship functions which are easier to handle from the mathematical or computational point of view without distorting or leaving information and giving thus the opportunity to simulate at least at a certain degree the "intuitive" behaviour of the human intelligence.

Records of a database, may be represented as points in a space, the position of the points being determined by the variables values which describes the records of the database. In principle the representation may also be reversed in the sense that the variables are represented as points in a space, while the position of each variable is defined by the records. This projection brings certain advantages. As a first technical advantage, certain relationship may be discovered which were hidden in the n-dimensional space of the information data being not intelligible either by human beings nor by machines, since the relative position of the records and/or of the variables in the space where the records or the variables are represented by points is a measure of their similarity or difference. A second technical

advantage is that the simplifying of the information data helps in transforming the data in data which may subjected to a computational evaluation and thus to help the machine to analyze the data to determine an

5   appropriate response to the data and to carrying out its computational job in a more rapid and simple way. One might not forget that for mathematical or computational problems there might be theoretically a solution which solution cannot be computed in practice.

10   The solution of a mapping problem allowing to reduce a three dimensional space for the data in a two dimensional space without loosing or distorting the information represented by the data has also a great relevance if one considers for example a machine which

15   collects image data from the environment and which has to generate an image recognizing the objects or at least discriminating certain objects between objects constituting obstacles and objects which do not constitute obstacles and also between objects that

20   might constitute obstacles at a later time. In this case a machine which has the possibility of reducing information about physical objects placed in a three dimensional space and which have a three dimensional extension in a two dimensional map would allow to

25   dramatically simplify the machine construction and to dramatically reduce the computational burden.

         The above described technical advantages are present already if one considers non humanoid machines having artificial intelligence. Considering for example

30   humanoid machines such like humanoid robots the advantages become more important since such a machine has a large number of sensors and a very high computing and evaluation burden is set to the processing units.

The algorithm to which the present invention relates has not only relevance for artificial intelligence, but can also help human intelligence in inspecting and analysing the relationships between
5   information data belonging to a n-dimensional space, where n is bigger then 3 by projecting the data onto a two or three dimensional space. This is a representation which can be understood by human intelligence having its senses constructed to sense a
10   three dimensional or two dimensional space. Thus a representation of data in this space can help human intelligence to understand and find out relationships which could be not be recognised in a four or more dimensional space.

15   Known algorithm for projecting data form a n-dimensional space into a less dimensional space, and particularly onto a three or two dimensional space uses a predetermined characteristic projection function for computing the position of each point in the projection
20   space. An example for such kind of projection algorithm is the so called Principal Component Analysis, briefly PCA which is described in H.Hotelling "analysis of a Complex of Statistical Variables into Principal Components" J. Educ. Psychol., 24:498-520, 1933 . This
25   algorithm provides the steps of defining N factors and N new variables which are orthogonal. Using this base of new variables a reorganisation of the data is carried out by attempting to put as much information as possible in the first factors under the constraint of
30   linearity. The mapping consist in rewriting the observations/variables using the computed factors and in plotting each one on a two dimensional map using as

coordinates the computed factors F1/F2, F3/F4 and so on.

This kind of projection algorithm working only on the base of linear projections determines that some information will be lost during the projection. In order to understand this situation consider a normal projection from a three dimensional space onto a two dimensional space. In a linear projections two points having a certain distance along one of the three dimensions might appear very near if the two dimensional projection space is perpendicular to the third dimension along which the two points are spaced apart. In a very simplified manner this situation takes place using PCA algorithm. The result of the known technique is that in the less dimensional space where the information data has been projected the data relationships is distorted in a dramatic way and the distortion can go so far as to cancel or abnormally enhance relationships between data.

The algorithm according to the present invention has the aim of projecting N-dimensional information data onto a less dimensional, particularly onto a two or three dimensional space without distorting in an excessive manner the relationships between data.

The algorithm according to the present invention has the following steps:

Providing a database of N-dimensional data in the form of records having a certain number of variables.

Defining a metric function for calculating a distance between each record of the database.

Calculating a matrix of distances between each record of the database by means of the metric function defined at the previous step

Defining a n-1 dimensional space in which each record is defined by n-1 coordinates.

Calculating the n-1 coordinates of each record in the n-1 dimensional space by means of an evolutionary
5 algorithm;

Defining as the best projection of the records onto the n-1 dimensional space the projection in which the distance matrix of the records in the n-1 dimensional space best fits or has minimum differences
10 with the distance matrix of the records calculated in the n-dimensional space.

As evolutionary algorithms so called genetic algorithms may be used.

Such kind of algorithm provide new solutions
15 basing on a starting parents population of solutions which may be computed according to various ways such as for example casual attempts. The solutions of the parent populations are combined in such a way that follows the basic combination of genes in genetics thus
20 giving new and different solutions which fitness score, for example in this case the error or difference form the distance matrices of the n- and the n-1 - dimensional spaces is evaluated for giving a certain relevance to the solution which will influence the
25 possibility of combination with other solutions of the new generation for generating a further generation.

This kind of computation makes use of an evolutionary algorithm in order to compute the position of the points in the projection space in such a way to
30 minimize the error with respect to the distances of the points in the original space and is always independent on the specific structure of the information data. Thus on the contrary to the PCA algorithm of the state of

the art the algorithm according to the present invention is does not use a predetermined characteristic projection function which computes the position of the points in the projection space.

The algorithm according to the present invention combines the projection of the information data with a particular evolutionary algorithm which will be described with greater detail in the following description of the examples.

More in detail and refered for simplicity to a projection into a two dimensional space, the mathematical problem which the present algorithm solves is the following:

Given N points and their distances in a L dimensional space, find into a 2 dimensional space the optimal distribution of these points according to the matrix of their constrained distances.

In strict mathematical language the above mentioned problem may be expressed as follows:

Defining a Map Distance in the two dimensional space such as for example:

$$Md_{ij} = \sqrt[2]{(Px_i - Px_j)^2 + (Py_i - Py_j)^2}$$

Where Md is the map distance and i and j are the number of the points and where PX and PY are the coordinates of the point in the two dimensional space.

Defining also a Vector Distance such as

$$Vd_{ij} = \sum_{k=1}^{L} \left| Pv_{ik} - Pv_{jk} \right|$$

Where Vd is the vector distance I and j are the indices of the different points and V_k are the vector components.

Thus the mathematical problem is to carry out the following optimisation:

$$\min E; \quad E = \frac{1}{C} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| Md_{ij} - Vd_{ij} \right|; \quad C = \frac{N \cdot (N-1)}{2}$$

Due to the reduction of the number of dimensions in the projection, there might be a situation in which two points might not be separated one from the other if the projection is carried out in a classical way. Thus no exact projection can be carried out form the mathematical point of view if information has to be not distorted or maintained at least partially in the less dimensional space.

The present algorithm solves the above problem by encoding each individual record represented by a point having coordinate X and Y. A set of different X and Y coordinates for each point is defined forming a first population of projections solution onto the less dimensional space, usually a two or three dimensional space.

For each of the projections of this first population the fitness score is calculated by using as the fitness function the matrix of distances of the single points in the originally N dimensional space. The population of projections is then subjected to combination according to the combination rules of the genetic algorithm thus producing a first generation population of projections which comprises X and Y coordinates for the points which are a combination of the coordinates provided in two projections of the parent generation.

The fitness score of the projections of the first generation is evaluated and again a new generation is formed basing on the first generation.

5   Using certain combinatory criteria of the projections of the parent generation based on the fitness score of this parent generation the genetic algorithm at each generation the genetic algorithm calculates solution having better fitness scores thus converging against the best solution.

10   Several genetic or evolutionary algorithm are known see for example, which differ one from the other mostly in the combinatory criteria of the parents in order to generate the next generation of solutions. This criteria relates to the admitted or forbidden

15   "marriages" of two individuals of the parent population and in the mechanism with which the two parents individuals combine their set of data, in this case the different coordinates of the points in the less dimensional map.

20   As an example a particular genetic algorithm used according to the invention is the so called Genetic Doping Algorithm disclosed in detail in BUSCEMA, 2000: M. Buscema, Genetic Doping Algorithm GenD), Edizioni Semeion, Technical Paper 22e, Rome 2000 and Massimo

25   Buscema & Semeion Group "Reti neurali artificiali e sistemi sociali complessi", Year 199, Edizioni Franco Angeli s.r.l. Milano, Italy, chapter 21, which disclosures are considered to be part of the present specification.

30   Briefly summarised the GenD algorithm provides for special modified rules for generating the new individuals of a following generation from the parents population.

As usual in the genetic algorithm, as a first step, GenD calculates the fitness score of each individual, depending on the function that requires optimisation, in this case the distribution function of

5   the data records in the general data set onto the training set and the testing set. The average health score of the entire population is then computed. Average health constitutes the criterion firstly of vulnerability, and secondly of recombination, of all

10   the individuals of the population, for each generation.

All individuals whose health is lower than or equal to the average health of the population are entered in a vulnerability list. This individuals are not eliminated, but continue to take part in the

15   process being only marked out. The number of vulnerable individuals automatically establishes the maximum number of marriages permitted for that generation. The number of possible marriages for each generation thus varies according to the average health of the

20   population. At the third step GenD algorithm couples the individuals. The entire population participate to this possibility. The maximum number of random coupling calls corresponds to half the number of individuals marked out as vulnerable.

25   For coupling purposes and the generation of children both the candidate individuals must have a fitness value close to the average fitness value of the entire population. Furthermore each couple of individuals may generate off-springs since it is

30   sufficient for marriage that at least one of the two individuals of the couple enjoy health values close to the health average of the entire population or even higher. According to another recombination rule GenD

algorithm does not consider possible marriages between two individuals of which one has a very low health value and the other a very high health value in comparison to the average health value of the

5   population. This means that too weak individuals and too healthy individuals tend not to marry themselves.

Recombination by coupling does not mean classic crossover of the genes of the parents individuals. GenD algorithm effects selective combination of the parents

10   genes by means of two types of recombination. A logic crossover; when repetitions are allowed and an opportunistic crossover; when repetitions are not allowed.

The logic crossover considers four cases:

15   1. Health of the father and mother are greater than average health of the entire population;

         2. Health of both parents is lower than the average health of the entire population;

         3. and 4. The Health of one of the parents is less

20   than the average health, while the health of the other of the parents is greater than the average health of the entire population.

If the case 1 does occur than recombination will be effected with a traditional crossover.

25   If the second case occurs, than the generation of the two children occurs through the rejection of parent's genes.

If case 3 or 4 occur, than the genes of the more healthy parent are transmitted to the children, while

30   the genes of the less healthy parent are rejected.

In the above the definition of rejection does not mean that the rejected genes are cancelled but only that these genes are substituted. Genes substitution is

not random but is carried out by means of a sliding
window criterion. Each gene may have different genetic
options or states. In this case substitution by a
sliding window means that the actual rejected gene will
be substituted by the same gene but having another
state as the original one. So during substitution the
criterion used by the GenD algorithm provide only the
substitution of the state of that gene which assumes a
different state as the gene had in the parent
individual.

Relating to the opportunistic crossover, this
crossover works when repetition are not allowed. In
this case the parents have overlapping genes with
respect to a random crossover point. In this case an
offspring is generated selecting the more effective
gene of the parents. The mechanism is repeated until
all the off-springs are completed.

A further criterion of the GenD algorithm rely
upon a final opportunity criterion which is a mechanism
that enables weak individuals being marked out and
having never had the opportunity to be part of a
marriage to re-enter the coupling mechanism thanks to a
mutation. The number of possible mutations is
calculated as the difference of the number of potential
marriages and the number of marriages carried out.
Mutations occur to those individuals which are present
and marked out in the vulnerability list. In this way
individuals which had never the opportunity to be part
of a generation process is given a final opportunity to
enter the evolutionary process.

From the above short explanation of the principal
features of this special genetic algorithm it appears
clearly that in the GenD algorithm the number of

marriages and of mutations are not external parameters, but adaptive self-definable internal variables, taking into account the global tendencies of the population system.

5   Furthermore It appears also clearly that the basic unit of the GenD algorithm is not the individual, unlike the classic Genetic Algorithm, but the species, which acts on the evolution of individuals in the form of the average health of the entire population of each

10   generation. The feedback loop between individuals and the average health of the population enables the present algorithm to transform in evolution terms the population as a whole from a list of individuals into a dynamic system of individuals.

15   As a further improvement step of the algorithm according to the present invention, a so called hidden point may be defined. This hidden point whose existence is only guessed is added in the parent population by giving to it position coordinates X_{hi} and Y_{hi} in the

20   projection.

The calculation of the evolutionary algorithm may be carried out in parallel with the hidden point and without the hidden point and the best fit projections obtained by the two parallel calculations may be

25   compared.   Hidden   points   might   help   in   better appreciating the peculiarities of the real positions of the points in the N-dimensional space and so better approximate this positions in the less dimensional projection.

30   Figure 1 reassumes briefly the mechanism of the present algorithm relatively to the evaluation with and without the hidden unit.

Although the example has been described in relation to a projection from a L dimensional space into a two dimensional space it is clear that the algorithm works similarly also fro projections in a
5   three dimensional space or in a L-1 dimensional space. Usually the projections in a two or three dimensional space are the preferred ones since the data representation can be better understood by human beings. When using the present invention for giving
10   more cognitive capacity to apparati, provided with artificial intelligence, also for dimensional or more dimensional spaces can be used if it is needed for carrying out the tasks for which the apparatus has been designed.

15   The projection carried out with the present algorithm may be on an Euclidean two or three dimensional space. Alternatively the way the algorithm calculate the projection might be understood as projecting the points for example not in a two
20   dimensional plane but on a two dimensional surface which is somehow curved and nevertheless represented graphically on a plane.

Given a certain database comprising a certain number of records each one characterised by a certain
25   number of variables, the present algorithm might be applied for projecting the database in two different ways.

A first way is to consider the records as being points and the variables as being the coordinates of
30   the points.

The second way is symmetrically reverse this situation by considering the records as variables.

The two spaces are defined as observation and variable spaces and the projections can bring to discovering relations between records and/or between variables.

5   In the following description of different examples it will be possible to appreciate the effectiveness of the present algorithm and the information which might be recovered by means of the hidden point and also by means of the two projections. From the specific examples it will also appear clearly the technical meaning of the present algorithm which goes beyond the fact of allowing to calculate in a very rapid way the less dimensional map of the points and in a fully independent way from the structure and meaning of the information data represented by each point or record. In the field of artificial intelligence and thus for example robotics this technical meaning resides in the fact that a computational machine may be able to analyse information data and to recognize or define relationships despite their complexity. The recognitions of relationships of information data in complex problem is important for giving to the machine not only computational power but also allowing the machine to take decisions relating the particular tasks the machine is destined to carry out.

The algorithm according to the present invention can be used for providing method for the cognitive analysis of multidimensional information data.

Provided a database comprising a certain number of records each one representing the relationship between one feature and a certain number of variables, a distance matrix of the records in the N-dimensional space defined by the number of variables characterising

each record is calculated according to a certain metric function.

This matrix is taken as the fitness matrix for the projection of the records or the variables into a less dimensional space, particularly a two or three dimensional space using the algorithm described above.

The representation of the records or of the variables in a two or three dimensional space may be used for recognising certain relationships among the records or among the variables.

An example of a particular problem to which the present method can be applied relates to the fact of determining the structure of a certain molecule when distances between at least some atoms forming the molecules are known. This problem is fo the kind so called not bound optimization problem.

The method can calculate a three or two dimensional projection of the structure of the molecule which may be graphically represented in an intelligible way for human beings. Further more relating to this problem when analysing complex molecules the projection can be carried out by adding the hidden individual which in this case could be an hidden atom and thus offering a tool for inspecting molecular composition and structures of highly complex molecules.

The method according to the present invention being independent on the structure of the information data can handle also not mathematical problems such as for example sociological problems. A first step of giving a numeric scale of evaluation of the different social variables considered might be provided in this case. Normally this is not a critical problem since such kind of variables often are characterised by

different state which can be defined by the value true, not true and not present so that the scale in this case could be defined as 1, (-1), and 0.

According to a further feature of the method the algorithm according to the present invention used for projecting the information data from the N-dimensional space onto a less dimensional space, particularly a two or three dimensional space may be applied in combination with other kind of projection algorithm which are somehow more sensitive to the information data structure.

A particular algorithm which can be used in combination with the present projection algorithm is the so called "SOM" Self Organizing Map algorithm which is a clustering algorithm The SOM is a known algorithm which is described in more details in KOHONEN, 1995: T. Kohonen, Self Organising Maps, Springer Verlag, Berlin, Heidelberg 1995 or Massimo Buscema & Semeion Group "Reti neurali artificiali e sistemi sociali complessi", Year 199, Edizioni Franco Angeli s.r.l. Milano, Italy, chapter 12.

SOM assumes a prior definition of the projection grid and projects codebooks of records in this grid via a competitive algorithm where dominant variables prevail over others.

The SOM projections gives rise to data clusters so called Kohonen units. The SOM algorithm is thus used to perform a first elaboration of the information data, while the algorithm according to the present invention is then used to reproject the Kohonen units emerging from the first elaboration in a coordinated and more detailed manner on its own map.

This procedure allows to take advantage of the peculiarity of the SOM algorithm to consider the significance of the variables and to take also advantage of the features of the present projection algorithm which can evaluate the fitness score of the projection performed referred to the fitness function which is the distance matrix of the points representing the information data I the N-dimensional space and which can also consider hidden units. Thus the reproduction accuracy in the less dimensional space is ensured and a more complex projection which can provide for greater information is performed. Figure 2 schematically illustrates the combination of SOM algorithm with the present algorithm. Thanks to the present algorithm the advantages of the SOM algorithm are combined with the fact that the present algorithm can dynamically deform the original projections space by hidden units increasing the reconstruction accuracy of the projection.

The invention relates also to a method for constructing two or three dimensional structural images of molecules starting from complete or incomplete data about the distances of the atoms of the molecule.

The invention relates also to a machine having artificial intelligence which is able to carry out actions based on processes simulating the "intuitive" reasoning of human intelligence when subjected to apparently not correlated data of non chemical or physical nature and/or collected physical or chemical data from the environment.

Figure 1 illustrate a diagram of the architecture of the projection algorithm according to the present invention.

Figure 2 illustrates a diagram of the combination of the algorithm according to the present invention with a SOM algorithm.

Figure 3 is a database of highway distances between Italian cities according to example I.

Figure 4 is the two dimensional projection map of the cities of the database according to the database of figure 3.

Figure 5 is a database of flight distances between US cities according to example II.

Figure 6 is the two dimensional projection map of the cities of the database according to the database of figure 5.

Figure 7 is a database of the European Countries Food Consumption in 1994 according to example III

Figure 8 is the two dimensional diagram of the projection of database of figure 7 in the variable space which has been elaborated by the algorithm according to the present invention.

Figure 9 is the two dimensional diagram of the projection of database of figure 7 in the observation space which has been elaborated by the algorithm according to the present invention.

Figure 10 is a table of 13 variables and their complement of a fourth example.

Figure 11 is the result of the projection of the variables according to the table of figure 10 on a two dimensional plane with the algorithm according to the invention.

Figure 12 illustrates the connection between the variables and the complements on the map according to figure 1.

Figure 13 illustrates a fifth method according to the present invention in which the projection algorithm is used in combination with a so called Self Organising Map.

Figure 14 is a further diagram illustrating the method according to figure 14 and the table of codebooks prototypes of the groups of variables defined by the projection of the database on the two dimensional map.

Figure 15 is a block diagram of an apparatus according to the invention having artificial intelligence.

Figure 16 is a graphic representation of a extremely simplified example of a situation where the apparatus having artificial intelligence, for example a robot, according to the invention can provide a correct response action.

Figure 17 to 19 illustrate an example of the method according to the present invention for generating two or three dimensional images of the structure of a molecule form incomplete data of the distance of the atoms forming the said molecule.

Example I

Example I clarifies the way the algorithm operates in order to generate a map from data relating to distances between objects.

In figures 3 and 4 there is described a first example of dataset and two dimensional mapping using the algorithm according to the present invention.

A dataset comprising ten Italian cities and their highway distances is provided. The highway distances are not true two dimensional distances in an Euclidean

Space, since every highway distance has three kind of alteration namely: a longitudinal alteration, an altitude alteration and a structural alteration. Thus creating a two dimensional map of the cities where the cities are placed considering only the highway distances using a linear algorithm would determine a distortion of the position of the cities with respect to their real relative position.

The city of Arezzo is not given to the algorithm with its distances from the other cities, but free determinable distances values are given to the algorithm so that the algorithm is called to look for a hidden city of which the existence is assumed and of witch the position is not known.

A first randomized distance value of the hidden city can be given in the distance matrix for the hidden city so that the algorithm can be initialised and can start to correct the randomized initial position of the hidden city. As it appears clearly from figure 4 a comparison of the map drawn by the algorithm according to the invention with a geographical map allows to identify the hidden city as the city of Arezzo.

Using the algorithm according to the present invention which carries out a non linear projection also the other cities being defined by their distances in the database are placed onto the two dimensional map by optimising their relative position with respect to the matrices of their relative distances. The distortion relative to the real position is very low and the solution is illustrated in figure 4.

EXAMPLE II

Example II is a similar mapping problem as example I. In this case the database comprises twelve US cities

and their relative flight distances. No hidden unit has been provided.

Also in this case the flight distances are affected by alterations similar to example I.

5   Also in this case a linear projection of the cities onto a two dimensional map would not take correctly care of the above mentioned alterations and the position of the cities on the map would be distorted relatively to reality.

10   The result obtained by the present algorithm is a map which is illustrated in figure 6 and where the positioning of the cities has an error of only 3,07% with respect to the matrix of distances while the positions of the cities is very close to the real

15   geographical position.

EXAMPLE III

Example III is more complex one. This example clarifies how the algorithm works in order to generate relations or correlations among data which apparently

20   have no logical relationship.

The database relates to the European Countries Food Consumption in nineteen ninety four. It comprises nine variables relating to the food kind, namely: cereals, rice, potatoes, sugar , vegetables, meat,

25   milk, butter, eggs.

Sixteen observations where made relating to sixteen countries, namely: Belgium, Denmark, Germany, Greece, Spain, France, Ireland, Italy, Netherlands, Portugal, Great Britain, Austria, Finland, Island,

30   Norway, Sweden.

The database was evaluated with the algorithm according to the present invention and the map according to figure 8 was obtained.

In the two dimensional map the circles indicates geographical areas to which the countries belong. The projection carried out by the present algorithm has shown that there are different groups of countries having similar food consumption and which countries belong to the same geographical area. Furthermore, the two dimensional projection has also highlighted that Ireland has a food consumption behaviour which is very different from that of all the other countries and particularly from the countries of the geographical area to which it belongs.

Figure 9 illustrates the projection of the database made by considering the records as variables, i.e. the observation countries as variables which has been defined as the observation space. The projection was also carried out by means of the algorithm according to the present invention and the map of figure 9 indicates also a relation which was not apparent from the database.

From the above it appears clearly that the algorithm according to the present invention carries out a projection which due to its non linearity does not lead to hidden information. The PCA algorithm needs to illustrate the information data onto two different maps for not loosing information, while the projection according to the present algorithm does not hide information and relationships between the data.

Entering more in detail, the degree of correlation between data is established by the present algorithm by means of a so called "share information" equation. This equation sizes the degree of association between two points in the map. This equation expresses the "share Information between two points representing two

different data records of a database in the original
multidimensional space in term of probability. Thus the
present method defines the degree of association
between two pints in the map as a probability of
5   association:

$$A_{i,j} = -\ln \frac{\dfrac{\sum\limits_{k=1}^{L}(1-Pv_{i,k})\cdot Pv_{j,k}}{L} \cdot \dfrac{\sum\limits_{k=1}^{L}(1-Pv_{j,k})\cdot Pv_{i,k}}{L}}{\dfrac{\sum\limits_{k=1}^{L}Pv_{j,k}\cdot Pv_{i,k}}{L} \cdot \dfrac{\sum\limits_{k=1}^{L}(1-Pv_{j,k})\cdot(1-Pv_{i,k})}{L}} \qquad Pv\in[0,1]; \quad A\in[-\infty,+\infty].$$

Returning to the capability of the present mapping
10 algorithm in finding out hidden units this capability
can be used for solving further technical problems as
for example for drawing two or three dimensional maps
of complex molecules also in the case where the list of
atoms is incomplete or where the matrix of the
15 distances is incomplete.

It has to be notices that as disclosed above, the
database can also be incomplete, this means that
despite knowing the presence of certain atoms the
distances of this ones may not be known. Thanks to the
20 ability of considering hidden units the algorithm
according to the present invention can place the known
atom of which the distances where not known in the
distance matrix in a correct or most probable position
relatively to the other atoms of the molecule.

25   According to another way of using the capability
of considering hidden units relating to this last
example, the algorithm according to the present

invention is also capable of considering the presence of unknown atoms in a molecule of which the composition is not completely known and further to this the algorithm is also capable of producing an hypothesis
5   about the most probable position of this atom relatively to the other known atoms thus helping the further study of the molecular structure.

    Example IV
10   In figure 10 the table of 13 variables and their complement is illustrated. The 13 variables relates to anagraphic data and medical data of a certain number of individuals, more precisely of 117 individuals. The aim is to analyse the database in order to find out
15 relations which are somehow connected to the Alzheimer disease or to the probability of developing the Alzheimer disease. Starting from the 13 variables the complement of this variables are defined. The complement being a complementary value of the
20 variables.

    Using the aforementioned database the data has been projected onto a two dimensional plane. The result is illustrated in figure 11. From this map the following conclusion can be drawn: The more two
25 variables are nearest, more their information is high and therefore the two variables are similar.

    In figure 12 the connection lines between each variable on the map and each complementary variable has been drawn in order to establish their relative
30 distance. From the mathematical point of view it can be demonstrated that the more the connecting segment is long the more the variable is significative in the database since his standardized variance is bigger.

EXAMPLE V

Figure 13 illustrates a diagram of a combined projection algorithm comprising two different algorithm one of which is a projection algorithm according to the present invention.

A database of different variables for a certain number of individuals comprises 19 variables of medical, anagraphical and social kind. The records of the database are elaborated with an algorithm known as Self Organising Map (SOM). This algorithm clusters the records into cells or units The database is an enlarge version of the one of Example IV.

The algorithm according to the present invention is applied to the units computed by the SOM in order to distribute the said units and the records clustered in it in an optimal way on a two dimensional map. Codebooks prototypes can be computed as the average of the codebooks of each unit taking part to a group.

The groups of units on the two dimensional map created by the projection algorithm are evaluated by means of their clustering on the map. It appears evident that the projection algorithm according to the invention will stress the existence of a fourth group.

Figure 14 illustrates a diagram in which starting form the database comprising the said 19 variables and subjecting the database to the SOM and afterwards to the projection algorithm four groups are generated on the projection map each group having is specific codebooks prototypes which are listed on the right table.

The variables considered are variables which can be involved in some way with the Alzheimer disease. The number of subjects considered has been of 117 patients. The different groups are characterised by different

5 percentage of patients having developed the Alzheimer disease. The codebooks prototypes can give insight in the relevance of certain medial variables and/or certain anagraphic variables and/or certain social variables for determining the risk of developing the

10 Alzheimer disease by an individuum.

It is interesting to notice that the age is not relevant while social variables such as intellectual level or level of schooling, physical exercise and other variables attaining to the behaviour has a high

15 influence in differentiating the four groups and thus the risk of developing Alzheimer disease. With increasing level of schooling and/o with increasing level of physical exercise and with increasing educational and cultural level the percentage of

20 individual having developed the disease becomes lower, despite the presence of certain pathological variables or medical variables which seems not to be relevant for differentiating the groups one from the other.

From the above mentioned projection different

25 suggestion may be extrapolated:

• Alzheimer disease at histological level starts independently from Tangles in Hippocampus or Plaques in NeoCortex, and arrives to Tangles in NeoCortex passing through Plaques in Hippocampus

30 with different transition probabilities. This suggestion is supported by evidences coming from the projection algorithm according to the present invention and SOM Systems.

- Severe Braak Stages are related to two different and unrelated pathologies (evidences supported by SOM System).

- Plaques in NeoCortex and Tangles in Hippocampus distribution are connected with two different kind of subjects in SOM System

- MMSE, ADL, BOSTON, and CNPR are strongly connected among them, in the same way that WRCL and VRBF are connected to each other. Evidences supported by the mapping through the projection algorithm according to the present invention which puts these two groups of tests in two different areas.

- Education Years are strongly connected with the Alzheimer disease pathology features (evidences supported by the algorithm according to the present invention).

- The integrated use of different Unsupervised Organisms, allows the identification of four natural clusters of subjects with specific codebooks prototype.

EXAMPLE VI

Example VI relates to a method for determining the conformation of a molecule of which the distance at least of certain of the atoms form at least other atoms forming the molecule are known.

In this case the method uses an algorithm according to the present invention for generating a map of the molecules which has the best fit in accordance to the known distances.

Thus a structure of the molecule can be drawn which does not loos, hide or distort information.

A database is formed which database comprises as variables the distances of the atoms from another atom of the molecule. The database is illustrated as a matrix in figure 17.

5   This database can be obtained by means of measurements carried out on the molecule as for example using radiographic inspection which is a common technique in Solid state physics for determining the lattice structure of cristals or other current

10   measurement means.

Once data has been achieved from the measurements, a database is generated in the form of a matrix were each row and each column having identical row and column index identify an atom. The distances between

15   each couple of atoms each one identified by row and column index are listed in the matrix and the result is a matrix where the diagonal elements have each one zero value and which is symmetric relatively to the said diagonal. Were no data about the relative distance

20   between two atoms is present a predefined value is given to the matrix elements. In the example the value (-1) has been chosen.

As a next step the algorithm according to the invention and to the previous description for

25   generating a map either in a two dimensional or in a three dimensional space is applied to the database. The results are shown in figures 18 e 19, respectively for the two and for the three dimensional map. In several experiments a fitness score and an error of the

30   generated maps relatively to the experimental data in the database of the measured distances has been calculated. Fitness scores in the range of 0,96 to 0,98 has been reached. Relating to the error, two kinds of

errors has been computed according to the below listed equations:

ERROR 1 being defined as

$$\sum_i abs(DistI_i - DistR_i)$$

5

and ERROR 2 being defined as

$$\sum_i \frac{abs(DistI_i - DistR_i)}{DistR_i}$$

were ERROR 1 is the sum of the absolute value of
10   the difference of the euclidean distances in the
computed map $DistI_j$ and the measured distances of the
database $DistR_j$ and were ERROR2 is the sum of the
absolute values of difference of the percentual error
of the euclidean distances in the computed map $DistI_j$
15   relatively to the measured distances of the database
$DistR_j$.

It has to be noticed that any variant of the
algorithm described above also in combination with
different examples can be used. In the present case the
20   drawing of two dimensional or three dimensional
molecular maps, these means the relative location as a
projection in a two dimensional or in a three
dimensional space, of the atoms of a molecule is a
problem which is analogous to the one of drawing a
25   geographical map described in the previous examples.

Starting from this reasoning a further improvement
of the method for determining the structure of a
molecule can also provide the steps of defining one or
more virtual or hypotetic atoms hereinafter indicated

as hidden atoms, which might exist in the structure of the molecule but have not been determined experimentally. This further step or improvement is analogous of the one of the hidden city in the mapping

5   example, were the city of AREZZO was introduced in the database without giving any distance value from other cities. Thus carrying out the method according to the present example, the algorithm will consider this one or more atoms and indicate them in the map. The hidden

10   atoms can be highlighted in the structural map of the molecule determined by the present method. As a result the method will provide for a prediction of the coordinates and/or of the distances of this one or more hidden atoms in the structure of the molecule form the

15   other atoms and the results can be used for deeper and more specific experimental inspections and structural analysis of the molecule which are aimed to verufy the real existence of this one or more predicted hidden atoms.

20   From this point of view the present method is thus alternatively also a method for carrying out an analysis of the structure of a molecule and for predicting the existence of further atoms in the molecule. Please consider that the virtual molecule of

25   the example has only 25 atoms which is a very low number of atoms if one considers organic chemistry or biochemistry were very big macromolecules are studied which have a much greater number of atoms.

        From a general point of view the method for

30   determining the structure of a molecule or a geographic map or for inspecting the structure of a molecule or a geographic map in order to investigate about the presence of a further "hidden" element and to predict

its position relatively to the other known elements of the structure of the molecule or of a map can be considered as a general method for generating maps and or a general method for predicting the existence of an hypotetical element of the map which is not apparent and which position relatively to the other elements is not known.

A further example of this method supporting its general principle is the generation of heaven maps or star maps and the prediction of the existence of a star or an other astronomical object and its position relatively to other astronomical objects basing on data relating to the relative distances of a certain number of stars or other astronomical objects. This kind of inspection avoid complex computations based on the observations, measurements and evaluations of orbital perturbations of objects and can give a first indication about the probability of the presence of an astronomical object which cannot be seen or is hidden.

The present invention relates also to an apparatus having artificial intelligence being capable of evaluating data which cannot be considered having a relation or correlation and for providing a reaction behaviour to environmental stimuli which is similar to the human intuition processes.

As already described, environmental data, may not show a direct or explicit correlation or relation. So these data and the corresponding reaction to them of a machine having artificial intelligence may not be computable or evaluated by the said machine causing the machine to enter a state of inactivity or blocking the machine. Nevertheless human beings are capable of evaluating these apparently not correlated stimuli by

determining in any case a reaction to them which might be either passive or active. If one considers machines having artificial intelligence such as robots or the like which has to interact with an ambient not being

5   created or adapted for them (for example by eliminating unnecessary stimuli for the tasks these machines has to carry out), then the skill of the robots has to increase dramatically in evaluating incoming stimuli which are disordered and which apparently have no

10   relation or no immediate recognizable relation among them and further to the evaluation in determining which reaction has to be carried out as a response to the collected stimuli.

     Thus a capacity of the machine or robot to

15   simulate a sort of "intuitive" behaviour of human beings can be of great technical importance.

     The apparatus according to the present invention is illustrated in figure 15 and comprises a processing unit 1. The processing unit can be similar to a

20   conventional computer for the parts relating to the electronics. The housing is formed in this case by at least part of the case or of the body of the apparatus. The processing unit is associated to a memory 2 for a program running the apparatus and which program is

25   carried out by the processing unit. A further memory 3 is provided for data which may be configuration data of the apparatus relating to input and output devices and to actuators or functional operating units or tools with which the apparatus is equipped and which are all

30   driven by the processing unit 1. The input devices can be of different kind and are summarised by the box 4 in figure 15. Input devices may be different relatively to the tasks the apparatus is designed for. So in a very

improved robotized apparatus such as an humanoid robot, the input device can be sensors interacting with physical and chemical stimuli such as any kind of mechanical, electric, acoustic electromagnetic or chemical stimulus. The number and kind of such sensors may vary depending on the conditions in which the apparatus has to operate. Other input devices can be provided such as input interfaces as a keyboard or reader for portable memory devices on which data is saved which cannot be sensed directly by the apparatus by means of its sensors.

The program saved in the program memory 2 has a plurality of routines or sections each one dedicated for carrying out a certain task and a plurality of routines for driving the actuators or operating devices or units and also routines for collecting the data acquired by the sensors and feeding the said data to an evaluation routine.

If one considers a highly improved apparatus having artificial intelligence, such as a robot which must have the capacity of carrying out basic functions simulating a human being, than considering the simple situation of such an apparatus passing in a street with a normal conditions of traffic it can be understood that the number of stimuli to which the apparatus is subjected are very high and that conditions may be present in which two stimuli of different kind having the same origin will arrive to the sensors of the apparatus in a condition in which the relation of having the same origin is not immediately or explicitly recognizable. Figure 16 try to graphically explain this condition graphically. Consider two vehicles A and B having a parallel path (indicated by the arrows 10, 11

considering a reflecting barrier on one side of the paths such as buildings 12, 13 or the like and considering also that a robotized apparatus 15 is located on the opposite side of the vehicles path as
5   referred to the reflecting barrier.

The apparatus has two visual sensors 115, such as two cameras and two acoustic sensors 215 such as microphones. This allow to collect three dimensional image information and three dimensional acoustic
10 information.

The acoustic waves generated and represented by the circles will arrive directly to the apparatus as acoustic stimuli indicated by the arrows 16 and 17. these stimuli will also arrive in the form of reflected
15 acoustic waves as indicated by the arrow 18, furthermore the direct and reflected acoustic waves of the two vehicles will also superpose or mix up.

In this condition the input data collected by the robotized apparatus 15 will provide certain number of
20 variables the relation of which is not immediately and directly apparent.

Thus forming a database in which the records consist in the collected input data and carrying out the mapping algorithm according to the present
25 invention as described before would provide a map in which the vicinity of the mapped data in the map would give to the apparatus a measure of the relation of the data one with respect to the other. Thus the apparatus is able to correctly relate the acoustic stimuli and
30 the visual stimuli by recognizing which sound is to be associated to the vehicle A and which sound is to be associated to the vehicle B. This is evaluation step provided by the algorithm according to the invention is

an analogous process as explained in the examples III, IV and V.

Furthermore considering the fact that reflected acoustic waves can be interpreted as a source of acoustic waves, by providing a hidden acoustic wave generator in the database, and using the algorithm of the present invention, the apparatus will be able to determine a map in which the reflectors are identified and their position and movement can be evaluated and also their relationship to the acoustic stimuli attaining to the reflected waves. Combining the mapping with the measure of relation, it could be also possible for the apparatus to identify to which one of the vehicles A or B the major component of the reflected wave has to be referred. Thus applying the method according to the present invention in the form of an algorithm which is coded in a program which is executable by a processing unit of an apparatus having artificial intelligence allows in the example described above to identify a relation between visual or image stimuli and acoustic stimuli, which relation was not directly recognizable from the stimuli sensed by the apparatus.

Although if constructing an appropriate model the above example could be solved by an equation describing the process based on the physical laws, it has to be understood that these way of solving the problem need at least an approximated a priori knowledge of the conditions of the ambient on which the model has to be constructed which is a very hard limitation. Furthermore in increasing the number of stimuli the problem might not be treated anymore by exact physical functions since with very large number of variables an

exact solvable computational problem may become not solvable. The apparatus according to the present invention comprising at least a processing unit 1 and a program memory 2 in which a program consisting in the present algorithm is saved for being carried out by the processing unit 1, so as to provide for a data processing method according to the present invention would in any case avoid an a priori knowledge of the conditions of the ambient in which the apparatus is located and would obviate to the necessity of generating exact mathematical models describing situations to which the apparatus may be subjected. Furthermore it would give the apparatus a possibility to operate also when it is subjected to a high number of stimuli.

The processing unit carries out the algorithm according to the present invention in the form of an executable program and generates the output map and the vicinity values as parameters related to the shared information as defined in the above description of the preceding examples. This out put is evaluated and used as input for decisional programs which drive the functions of the actuators of the apparatus. It is also possible to provide learning programs for saving the event as experience of the apparatus, thus refining the further computational tasks.